基于 HPSG 的汉语词库和语法规则系统构建*

杨春雷
(上海外国语大学英语学院 上海 201600)

摘要:【目的】探讨开发汉语可计算语法的理论基础和实践经验。【方法】以 HPSG 理论为句法框架, 以 MRS 为语义描写手段, 以"汉语普通话在线语法"(简称"汉构")的开发过程为例, 重点研究通过构建词库和规则层级, 对汉语特殊结构进行计算实现。【结果】"汉构"的开发证明, HPSG 非常适合作为汉语可计算语法开发的理论框架。【局限】"汉构"仍在继续开发中, 其覆盖率暂未接受大规模自然语料的检测。【结论】"汉构"可作为沟通形式语言学和计算语言学的桥梁和开发大规模资源型语法的基础。

关键词: HPSG 汉语普通话在线语法(汉构) 语法工程 自然语言处理
分类号: H087 G350

1 引 言

自然语言处理(Natural Language Processing, NLP)方法可分为浅层和深层处理。前者指基于数据和统计的处理方法, 在 20 世纪 90 年代成为 NLP 的主要方法[1], 并在语音、拼写和语法检查等领域促成了一些重要研究成果。但是, 由于自然语言非常复杂, 计算机在执行复杂的分析任务时, 速度缓慢, 空间不足, 无法投入处理效率要求较高的商业应用。因此, 计算语言学家意识到要提高 NLP 的精确度, 并降低对计算系统的要求, NLP 必须回到基于精确的语言学模型的方法, 即基于约束条件或规则的深层处理的方法。其中, 编写机读的形式语法的程序[2], 即可计算语法开发或语法工程(Grammar Engineering)是关键。可计算语法开发过程复杂, 要建立词库、短语和句法规则、语义表达、信息结构等不同层次且相互关联的部分。从 20 世纪末至今, 面向深层语言处理的可计算语法开发经历了近 20 年平稳快速发展期, 并在语言学理论基础、跨语言、计算实现技术和商用等方面取得了重要进展[3-5]。

但是, 汉语 NLP 仍落后于英语、日语、德语等语种。尽管学界已开发出多种分析汉语形态特征的分析器, 但尚未有一部系统地面向深层语言处理的可计算语法。一方面, 计算语言学家对汉语的复杂结构和语义特征感到很棘手, 无法取得高效的分析结果; 另一方面, 汉语语法学家对计算语言学中应用的语言学框架、可计算语法开发平台和工具缺乏了解, 为计算语言学提供的可借鉴成果有限。陆俭明[6]认为中文信息处理"眼下特别要加强词汇句法语义研究, 集中精力解决好'句处理'问题", 并一针见血地指出"语言研究已成为信息工程科学发展的瓶颈"。虽然 NLP 是一个多边缘的交叉学科, 但应以语言学为主[7]。

在 HPSG 理论框架内, 根据相似的编写流程, 国际上已经开发出 9 种语言的大规模语法, 并且已投入商用, 另外, 还有其他 10 余种语法正在开发和完善(http://wiki.delph-in.net/moin/GrammarCatalogue)。笔者和斯坦福大学语言与信息研究中心共同开发"汉语普

通讯作者: 杨春雷, ORCID: 0000-0001-9123-7502, E-mail: yangchunlei@shisu.edu.cn。

*本文系国家社会科学基金规划一般项目"类型学视野下的汉语短语结构语法及其计算现实研究"(项目编号:16BYY136)、教育部人文社会科学研究规划基金项目"面向深层语言处理的汉语短语结构语法"(项目编号:13YJC740118)和上海外国语大学规划基金项目"语言量化现象的多维度研究"(项目编号:2013XJGH023)的研究成果之一。

通话在线语法”(Mandarin Grammar Online, ManGO 或“汉构”)[8]是最早开发出的汉语可计算语法之一。本文结合汉构的开发实践，讨论开发汉语可计算语法的理论基础、技术思路和主要开发环节(构建词库和规则系统)。

2 汉构的理论基础和开发过程

2.1 汉构的理论基础

句法方面，汉构基于中心语驱动的短语结构语法(Head-driven Phrase Structure Grammar, HPSG)[9-12]。HPSG 使用特征结构的类别层级(Type Hierarchy)构建各层次的语符，使用约束条件规定语符的合法性。使用 HPSG 能够系统高效地构建可计算语法，同时保持理论语法精准的形式表达和扎实的理论基础。HPSG 理论句法语义并重，借鉴了许多语言学理论的描写手段和研究成果，非常适合开发跨语言且覆盖广泛的语法体系，其主要特点是非转换、基于约束条件、表层导向、高度词汇化等[13]。这些特点非常适合汉语可计算语法开发。因此，近十几年，许多汉语语法学家一直呼吁重视 HPSG 理论对汉语语言学研究的特殊重要作用。由于 HPSG“不仅具有较广泛的描写语言现象的能力，而且所做的描写也比较自然”[14]，而汉语中丰富的词汇特征在很大程度上决定了句法和语义结构，所以 HPSG 的词汇主义特征特别适合汉语分析[15]。此外，根据欧洲专家咨询小组(European Expert Advisory Group)发布的报告，HPSG是计算语言学领域应用最广泛的语法理论[16]。

语义方面，汉构使用最小递归语义(Minimal Recursion Semantics, MRS)描写语义[17]。MRS 采用平型(Flat)语义形式表达系统，提供量化词和辖域算子(Operator)的不详描写(Underspecification)，可以在句法没有确切描写语义约束条件时，对各层次的语义约束条件进行编码，同时不会增加错误的句法歧义。MRS 适用于基于分类特征结构(Typed Feature Structure)的自动剖析及生成语句，已经在基于 HPSG 的多语种计算语法开发、研究、教学和商用实践中被证明非常灵活高效[18]，同样适用于汉语可计算语法开发[19]。

汉构的开发和测试平台是“语言知识建构系统(Linguistic Knowledge Building system, LKB)”①。LKB 是专为基于约束条件的语言学形式体系(如 HPSG)设计的词库和语法开发平台。

2.2 汉构的开发过程

汉构的 HPSG 属性决定了它没有任何派生或转换性质的操作，只包括具体的句子成分结构、一组数量有限的语法规则、普遍原则和富含语法及语义信息的词库。汉构的编写过程包括定制语法、建立测试套件、建设词库、描写语法规则等环节[8]。

汉构的定制语法来自“语法母体(Grammar Matrix)”②。该语法项目的目的是建立一个不同语法共享的内核，主要包括基本特征结构、技术手段、匹配语义描写的类别、基本规则与结构类别等信息[4]。开发者可以通过参数化设置，自动生成针对目标语言的语法现象的形式化描写，作为初始语法。从 2001 年起，该项目组根据普遍语法特征研究，致力于建立跨语言的可计算语法的基础，迄今共开发了 20 多种基于语法母体的可计算语法。

汉构面向和使用的是 MRS 测试套件③。该测试套件有 12 种语言的平行语料，覆盖丰富且具代表性的语言现象，已用于多种计算语法的开发。

但是，初始语法远未达到令人满意的精确度和覆盖面。仅以 MRS 测试套件为例，在构建第一版词库后，初始语法只能自动剖析不到三分之一的例句。这说明，开发者还需要针对汉语特点，构建更精确的词库和语法规则系统。这两个环节相对较为复杂和关键，只能由开发人员在初始语法的基础上，通过人工继续拓展和完善。与正在开发的其他汉语可计算语法相比，汉构的特点在于挑战汉语比较特殊的语法现象[20]。笔者结合汉构开发的具体实践，针对汉语词汇特点和特殊的语法现象和结构，重点讨论如何通过构建汉语词库和语法规则系统，提高计算语法的精确度和覆盖面。

①http://moin.delph-in.net/LkbTop.

②http://www.delph-in.net/matrix/.

③http://moin.delp-in.net/MatrixMrsTestSuite.

3　构建词库：词项和类别层级

汉构的词项由类别定义、书写形式和语义信息三部分构成，以“追赶”为例：

```
追赶_v := v_trans-verb-lex &          ;; 类别定义
  [ STEM < "追赶" >,                  ;; 书写形式
    SYNSEM.LKEYS.KEYREL.PRED
      "_zhui1gan3_v_rel" ].           ;; 语义信息
```

词项和语法规则的形式化描写包括结构类别定义和特征描写两部分，由符号&连接。前者由符号:=引入更高层级的结构定义，读为“属于”；后者使用方括号[ ]内的特征结构描写。大写字母表示 HPSG 术语，如 SYNSEM 表示句法语义联合体，LKEYS 表示词项的语义指针；KEYREL 表示关键关系指针；PRED 表示关系的谓词名称。小写字母表示句法概念，如 v_trans 表示及物动词，verb 表示动词，lex 表示词项等。双分号是对形式化描写的文字说明。该词项第一行规定“追赶”属于及物动词类别(即 v_trans-verb-lex)；第二行是书写信息；第三行是语义信息。[ ]内特征之间的句号“.”表示特征结构的路径，自左至右层级越来越低，靠左的特征结构包含靠右的结构。如[KEYREL.PRED]表示位于更高层级的关键关系特征(KEYREL)包含谓词名称特征(PRED)。

通过相关联的词汇层级、曲折变化规则、语法规则和语义关系，词项逐级投射到合法的句法和语义结构中，逐渐形成完整的句法结构分类层级。例如，及物动词“追赶“的词汇类别层级由上而下如图 1 所示：

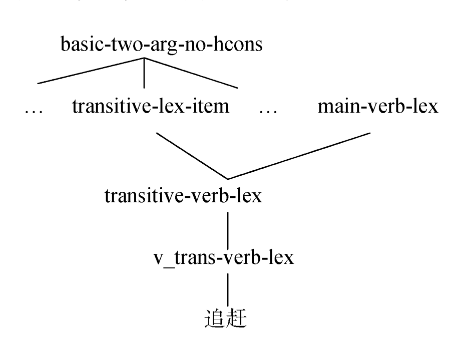


图 1　及物动词“追赶”的类别层级

层级越高的结构描写越概括，层级越低描写越具体，并继承所有母节点的特征描写。在下面的描写中，ARG-ST 表示论元序列，LOCAL 表示本地性，CAT 表示范畴，VAL 表示配价，SPR 表示先行语，COMPS 表示补语。符号< >表示序列，例如[ARG-ST <ARG1, ARG2, ARG3>]表示论元序列包含三个论元。需要特别说明的是，HPSG 是表层导向的语法理论，其描写与表层语言结构的语序严格对应。因此，一般情况下，汉语的 ARG-ST 序列中第一个论元对应主语，第二个对应宾语。

匹配句法和语义信息，即在第二个论元的语义指针和该动词所在的二元谓词结构中相应论元角色(即 ARG2)之间建立联系。此外，因为汉语是主语脱落型的语言，描写还规定第一个论元可以为空。

```
transitive-lex-item := basic-two-arg-no-hcons &
  [ ARG-ST < [ ],     ;; 第一个论元可为空
        [ LOCAL.CONT.HOOK.INDEX ref-ind & #ind2 ] >,
         ;; 第二个论元的语义指针
    SYNSEM.LKEYS.KEYREL.ARG2 #ind2 ].
      ;; ARG2 和语义指针共指
```

需要进一步规定及物动词涉及的两个论元的详细特征。在下面的 transitive-verb-lex 词汇类别定义中，ARG-ST 特征的两个序列值规定它们的中心语都是名词。该定义还再次匹配了句法和语义信息，明确论元结构中的第一个成分(通常是主语)的语义指针指向该动词所在的二元谓词结构的相应论元角色(即 ARG1)。

```
transitive-verb-lex := main-verb-lex & transitive-lex-item &
  [ SYNSEM [ LOCAL.CAT.VAL.COMPS < #comp >,
      LKEYS.KEYREL.ARG1 #index ],
    ARG-ST < [ LOCAL [ CAT.HEAD noun,
        ;; 第一个论元的中心语是名词
              CONT.HOOK.INDEX #index ] ],
                ;; ARG1 和语义指针共指
              #comp &
              [ LOCAL.CAT [ VAL [ SPR <>,
                                  COMPS <> ],
                      HEAD noun ] ] > ].
                      ;; 第二个论元的中心语是名词
v_trans-verb-lex := transitive-verb-lex
```

HPSG 是高度词汇化的语法理论，词库中的词项蕴含丰富的句法信息。与英语、德语、西班牙语等许多语言相比，汉语缺少形态和句法标记。因此，汉语 NLP 可借助的语法手段很有限。例如，汉语中没有一致和格的概念，曲折变化形式也很少，时体通常由非常有限的词汇手段表示，如“了”、“着”、“过”和时间状语等。但是，汉语的词汇信息非常丰富，存在很多兼类词，且影响句法和语义结构。因此，词库构建和词项定义对汉语 NLP 尤为重要。在汉语词库构建中，词项的定义往往并非取决于词形，而是取决于具有不同句法功能的义项。LKB 在剖析句子时，只有当词库中没有任何对应字符串的情况下才会给出词汇项缺失提示。

如果分析失败，只有通过分析树图才能找出与词项定义相关的原因，显然那样做会更耗时费力。建立词库时，要根据所描写的语法现象充分考虑特定词形的多重句法功能，完善词项定义，否则会引起剖析失败。如例 1 中的“给”具有多种句法功能：

例 1: (a) 张三给了李四。(二元及物动词)

(b) 张三给李四书。(三元物动词)

(c) 张三给李四打了。(复杂谓语)

(d) 张三把书拿给李四。(“把”字句)

(e) 张三给李四拿书。(介词)

这些句法功能决定了需要在词库中添加哪些词项。例如，“给”的 4 个词项定义如下所示：

```
给_v:= v_trans-verb-lex &     ;; 二元动词，用来描写(a)类型的句子
  [ STEM < "递" >,
    SYNSEM.LKEYS.KEYREL.PRED "_di4_v_rel" ].
给_v1 := ditrans-verb-lex &   ;; 三元动词，用来描写(b)类型的句子
  [ STEM < "给" >,
    SYNSEM.LKEYS.KEYREL.PRED "_gei3_v_rel" ].
给_v2 := v_light-verb-lex &   ;; 轻动词，用来描写(c)类型的句子
  [ STEM < "给" >,            ;; “给”相当于“被”
    SYNSEM.LKEYS.KEYREL.PRED "_gei3_v_lt_rel" ].
给_p := prep-no-mod-lex &     ;; 介词，用来描写(d-e)类型的句子
  [ STEM < "给" >,
    SYNSEM.LKEYS.KEYREL.PRED "_gei3_p_rel" ].
```

由于汉构的词库是面向 MRS 测试套件而建，仅提供套件中出现的词形所对应的词项，因此词项的数量并不大，但由于每个词项的句法信息丰富，因此构成的词项类别比较丰富。据统计，汉构词库共有192个词形，231个词项，76个词项类别。

4 构建语法规则系统

建好词库后，还需要定义和描写各种语法规则并构建类别层级，才能把词项组合成更大单位的合法结构。汉构的规则系统包括词汇规则、短语规则、句法规则和原则 4 个部分。其中，原则本质上也是规则的一种，包括少数约束全部句法结构的核心规则；词汇规则主要用来生成各种曲折变化形式；短语规则用来生成短语结构；句法规则用来生成更复杂的小句结构。汉构的规则系统如图 2 所示。

以句法规则为例，为构成小句，需要一些基本的句法规则，例如“主语–中心语”规则和“中心语–补语”规则，分别如图 3 和图 4 所示。

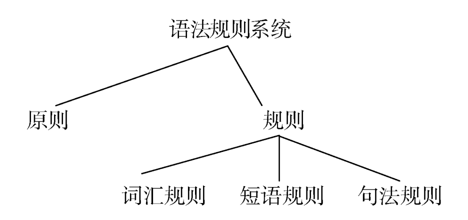


图 2 汉构的语法规则系统

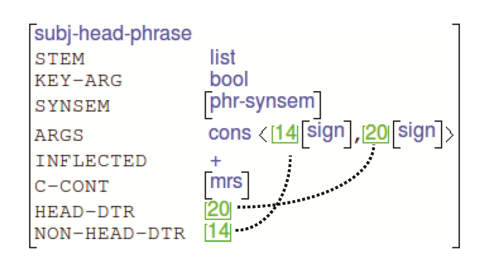


图 3 “主语–中心语”规则的特征结构

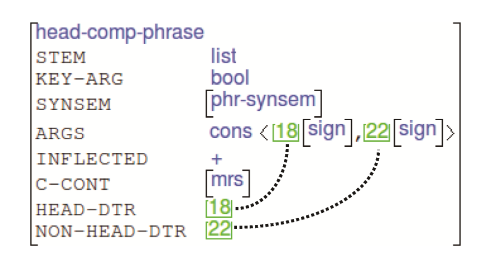


图 4 “中心语–补语”规则的特征结构

在图 3 中，第[14]项是主语，即非中心语成分(NON-HEAD-DTR)；第[20]项是中心语节点。而且，在表层结构 ARGS 中，[14]位于[20]之前。同理，图 4 中的“中心语–补语”规则规定中心语节点[18]位于非中心语节点，即其补语[22]之前。

再描写中心语、主语和补语的特征。例如，规定汉语的中心语包括动词(如“追赶”、“认为”和“叫”)、形容词、名词(如“张三”、“李四”)和介词；主语是名词性成分；补语包括名词性成分或小句(如“李四在叫”)。这两条基本的句法规则可以构成 SVO(主语+动词+宾语)结构(如例 2 和图 5)和 SVC(主语+动词+补语)结构(如例 3 和图 6)。

例 2: 张三追赶李四。

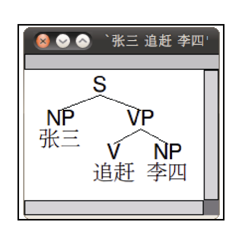


图 5 SVO 结构“张三追赶李四”的树形图

例 3：张三认为李四在叫。

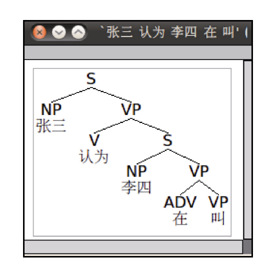


图 6　SVC 结构“张三认为李四在叫”的树形图

将每条句法规则直接和子节点的关系序列联系在一起，构成这些句子的 MRS 语义表达。该过程从最底层由词项实现的关系开始，通常每个词项表示一种关系。句法规则也可能直接把语义内容整合在一起，但这里讨论的基本句法规则只是将子节点上的词和短语的语义内容集合起来。语义组合方面的确切约束条件通过丰富的词汇类别定义实现，如之前讨论过的及物动词类别定义和层级。

但是，仅依靠基本句法规则远远不够，为了能自动剖析汉语特殊结构，需要在定制语法的基础上增加有针对性的规则。例如，定制语法无法分析例 4，其剖析流程如图 7 所示。

例 4：追赶猫很无聊。

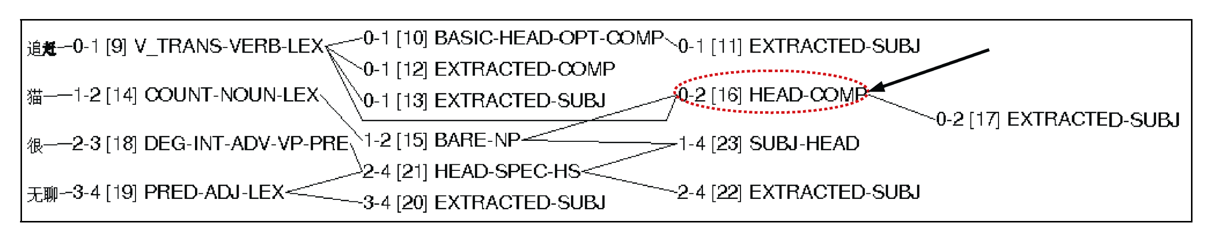


图 7　“追赶猫很无聊”的自动剖析流程图

通过流程图可以看出词项类别、应用的短语和句法规则、句法组合步骤等详细信息。流程图的每个节点包括三部分信息。以图 7 中箭头所指的被虚线圈出的部分为例，第一部分用两个阿拉伯数字标明的区间(即 0-2)，表示该节点覆盖的词项范围。词项标注按照 0-1(第一个词项)、1-2(第二个词项)、2-3(第三个词项)的顺序依次进行。0-2 表示此节点覆盖了前两个词项。第二部分是方括号中的数字(即[16])，表示句法组合的步骤。最后一部分大写字母是 HPSG 的术语组合，表示在该步骤形成的句法结构。“0-2 [16] HEAD-COMP”表示在第 16 步组合形成“中心语–补语”结构，涵盖了(0-1) “追赶”和(1-2) “猫”两个词项。

图 7 显示虽然通过“中心语–补语”规则形成了 VP “追赶猫”，但没有句法规则允许它做主语，导致其无法进一步与“很无聊”组合。汉语中 VP 做小句主语的情况很常见。在英语这种曲折变化形式较丰富的语言中，非谓语动词形式，如动名词，可以帮助甄别含有动词形式的主语，如例 5，其自动剖析结果如图 8 所示。

例 5: Chasing the cat is boring.

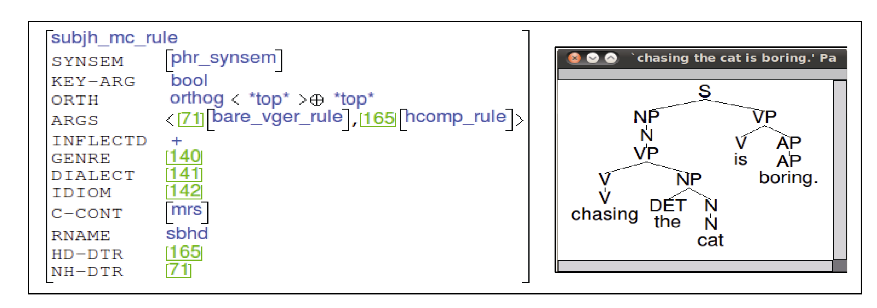


图 8　Chasing the cat is boring 的特征结构和树形图

其中，chasing 由“动词原形 chase + -ing 标记”构成，可以通过曲折变化短语规则把带有“-ing”标记的动词短语转换为名词性成分。但汉语中缺少动名词曲折变化形式，导致定制语法无法分析例 4。为解决这种汉语特殊现象，需要增加允许 VP 做主语的规则。在汉构中，定义汉语的主语为包含可选择论元的 olist 序列，即“追赶”的主语可缺省。因此，VP “追赶猫”就可成为饱和的(Saturated)小句成分结构，这意味着它可以担

当后面的谓词成分“很无聊”的主语，从而能够根据“主语–中心语(SUBJ-HEAD)”规则进行下一步组合。相关具体描写如下所示。汉构最终成功剖析例 4 的结果如图 9 所示。

```
olist := list.
subj-head-phrase := decl-head-subj-phrase & head-final &
  [ HEAD-DTR.SYNSEM.LOCAL.CAT [ VAL [ SPR olist,
                                  COMPS <> ],
                            POSTHEAD + ] ].
```

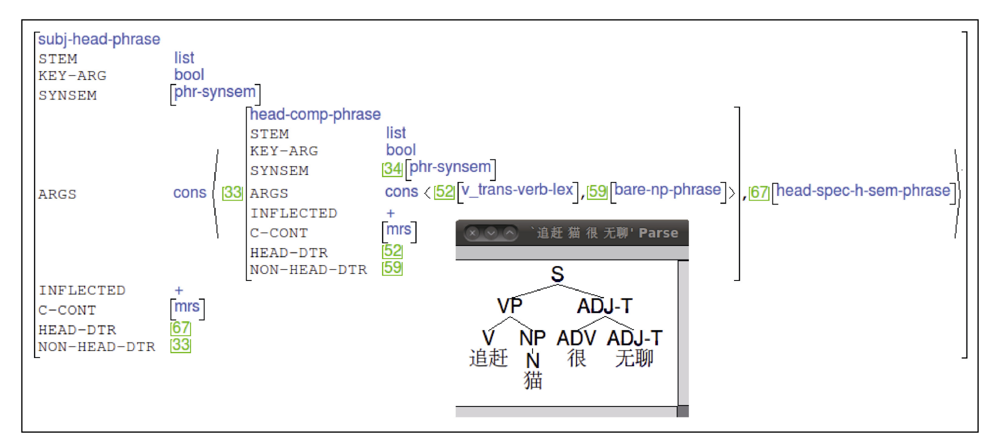


图 9　“追赶猫很无聊”的特征结构和树形图

汉构编制了许多针对汉语特殊语法现象和结构的规则，因此有效覆盖了较为广泛的汉语自然语料。例如，成功处理例 4 应用了一条特殊的“表语形容词转换”词汇规则，规定汉语形容词不需要系动词(如例 5 中的 is)就可直接做谓词。此外，通过定义汉语主语的序列类别，可以在保持句法规则简洁的同时避免词库信息冗余，从而保持整个系统的俭省性。例如，不必把例 4 中的“追赶”分别列为名词和动词两个词项。相反，如果在词库中将所有兼类词单列为词项，由于汉语兼类词非常普遍，词库规模会成倍扩大。

5　结　语

汉构的开发至今已持续约 7 年，其中集中开发约两年，而且还在不断更新。汉构的成功开发，证明了 HPSG 框架内的汉语语言学理论分析和计算实现不仅可行，而且高效。它在形式语法理论和计算语言学之间搭建起一个有效的跨学科平台。汉构已经成为一些最新开发的计算语法的基础，例如新加坡国立大学系统科学研究院组织开发的 ZHONG[|]语法系统[21-22]。目前，汉构的语法体系共有大约 5 200 行语法规则描写，涵盖了相当广泛的汉语现象，包括时体貌、属格形式、介词性修饰成分、结果状语结构、并列结构、内嵌小句、名物化、“被”字句、“把”字句和兼语式[23]等。根据语法性能测试工具[incr tsdb()](TSDB)的批处理理结果，汉构已经完全覆盖了汉语 MRS 测试套件[8]。

此外，汉构使用 MRS 为每个句法自动剖析的结果匹配了组合而成的语义分析。因此，它具有完全可逆的特点，即除了用于传统的句法自动剖析，还可根据 MRS 输入自动生成合法的汉语句子。

下一步，笔者计划加强针对汉语特殊结构的语言学本体研究，并将研究结果计算实现，拓展汉构的覆盖面，提高处理效率。具体地，需要建立一个包含更丰富语言现象的测试套件，扩建词库，并通过修改完善各级规则，以及检查使用语法自动生成的不合理语句，减少错误的剖析结果。

(致谢：感谢匿名外审专家以及编辑部的修改意见。)

参考文献：

[1] Hutchins J. Latest Developments in Machine Translation Technology: Beginning a New Era in MT Research [C]. In: Proceedings of the International Cooperation for Global Communication, Kobe, Japan. Tokyo: AAMT, 1993: 11-34.

[2] Bender E M, Flickinger D, Oepen S. Grammar Engineering for Linguistic Hypothesis Testing[C]. In: Proceedings of the Texas Linguistics Society X Conference: Computational Linguistics for Less-studied Languages. 2008: 16-36.

[3] Oepen S, Flickinger D, Tsujii J, et al. Collaborative Language Engineering: A Case Study in Efficient Grammar-based Processing [M]. CSLI Publications, 2002.

[4] Bender E M, Drellishak S, Fokkens A, et al. Grammar Customization [J]. Research on Language and Computation, 2010, 8(1): 23-72.

[5] Bender E M. Linguistic Fundamentals for Natural Language Processing: 100 Essentials from Morphology and Syntax [J]. Synthesis Lectures on Human Language Technologies, 2013, 6(3): 1-184.

[6] 陆俭明. 汉语言文字应用面面观[J]. 语言文字应用, 2000(2): 4-8. (Lu Jianming. Aspects of Language Use in China [J]. Applied Linguistics, 2000(2): 4-8.)

[7] 冯志伟. 自然语言处理的学科定位[J]. 解放军外国语学院学报, 2005, 28(3): 1-8. (Feng Zhiwei. Academic Position of Natural Language Processing [J]. Journal of PLA University of Foreign Languages, 2005, 28(3): 1-8.)

[8] 杨春雷, Dan Flickinger. 汉构: 面向深层语言处理的语法工程[J]. 现代图书情报技术, 2014(3): 57-64. (Yang Chunlei, Dan Flickinger. ManGo: Grammar Engineering for Deep Linguistic Processing [J]. New Technology of Library and Information Service, 2014(3): 57-64.)

[9] Pollard C, Sag I. Information-based Syntax and Semantics, vol. 1: Fundamentals [M]. CSLI Publications, 1987.

[10] Pollard C, Sag I. Head-driven Phrase Structure Grammar [M]. University of Chicago Press, 1994.

[11] Sag I A, Wasow T, Bender E M. Syntactic Theory: A Formal Introduction [M]. CSLI Publications, 2003.

[12] Boas H C, Sag I. Sign-based Construction Grammar [M]. CSLI Publications, 2012.

[13] Kim J B. The Grammar of Negation: A Constraint-based Approach[M]. CSLI Publications, 2000.

[14] 方立, 吴平. 中心语驱动短语结构语法评介[J]. 语言教学与研究, 2003(5): 31-43. (Fang Li, Wu Ping. A Review of Head-driven Phrase Structure Grammar [J]. Language Teaching and Linguistic Studies, 2003(5): 31-43.)

[15] 陆俭明. 句法语义接口问题[J]. 外国语, 2006(3): 30-35. (Lu Jianming. On Interface Between Syntax and Semantics [J]. Journal of Foreign Languages, 2006(3): 30-35.)

[16] Backofen R, Becker T, Calder J, et al. The EAGLES Formalisms Working Group-Final Report [C]. In: Suresh Manandhar, Anne-Marie Mineur, and Gertjan. 1996.

[17] 曾少勤, 王惠临, 张寅生.汉语文本的最小递归语义表示研究——以名词性量化短语为例[J]. 现代图书情报技术, 2012(10): 35-41. (Zeng Shaoqin, Wang Huilin, Zhang Yinsheng. Mandarin Text Representation Based on Minimal Recursion Semantics – Illustrated by Quantitative Noun Phrases [J]. New Technology of Library and Information Service, 2012(10): 35-41.)

[18] Copestake A, Flickinger D, Pollard C, et al. Minimal Recursion Semantics: An Introduction [J]. Research on Language and Computation, 2005, 3(2): 281-332.

[19] Kim J B, Yang J. Korean Phrase Structure Grammar and Its Implementations into the LKB System [C]. In: Proceedings of the 17th Pacific Asia Conference on Language, Information, and Computation. 2003: 88-97.

[20] Fokkens A, Avgustinova T, Zhang Y. CLIMB Grammars: Three Projects Using Metagrammar Engineering [C]. In: Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12), Instanbul, Turkey. 2012: 1672-1679.

[21] Fan Z, Song S, Bond F. An HPSG-based Shared-grammar for the Chinese Languages: Zhong[|] [C]. In: Proceedings of the Grammar Engineering across Frameworks (GEAF) 2015 Workshop. 2015: 17-24.

[22] Fan Z, Song S, Bond F. Building Zhong, a Chinese HPSG Shared-Grammar [C]. In: Proceedings of the 22nd International Conference on Head-Driven Phrase Structure Grammar. 2015: 96-109.

[23] 杨春雷. 兼语式的深层语言处理: 从语言学设计到计算实现[J]. 外国语: 上海外国语大学学报, 2013, 36(3): 50-59. (Yang Chunlei. Deep Linguistic Processing of Pivotal Construction: From Linguistic Design to Implementation [J]. Journal of Foreign Languages, 2013, 36(3): 50-59.)

利益冲突声明:

作者声明不存在利益冲突关系。

支撑数据:

支撑数据由作者自存储, E-mail: yangchunlei@shisu.edu.cn。

[1] 杨春雷. MRS-utf8-Full-annotated2 Justin__April 22.UTF8. MRS 测试套件.

[2] 杨春雷. lexicon.tdl. 词库文件.

[3] 杨春雷. Flickinger D. mandarin.tdl. 语法规则系统.

收稿日期: 2016-01-25
收修改稿日期: 2016-03-18

Building Online System for Chinese Lexicon and Grammar

Yang Chunlei
(College of English Studies, Shanghai International Studies University, Shanghai 201600, China)

Abstract: [**Objective**] This paper explores the theoretical foundation and practical experience of buidling a computational Chinese grammar system. [**Methods**] This study discussed the development process of the Mandarin Grammar Online (ManGO), an Head-driven Phrase Structure Grammar (HPSG) system with Minimal Recursion Semantics. It built the lexicon and hierarchy rules for the idiosyncratic structures of the Chinese grammar. [**Results**] The successful development of the ManGO system showed that the HPSG was an ideal theoretical framework for the Chinese computational grammar applications. [**Limitations**] ManGO was still underdeveloped, and it was not able to examine this system's coverage with large-scale natural language data. [**Conclusions**] ManGO connects the theories of formal and computational linguistics, therefore, it becomes the foundation to develop large scale resource grammar.

Keywords: HPSG Mandarin Grammar Online (ManGO) Grammar engineering Natural Language Processing

---